Deep-Learning Based Immunohistochemistry Scoring Predicts Progression and Prognosis of Human Esophageal Cancer

Abstract. Esophageal cancer (EC) is a highly lethal malignancy worldwide with a 5-year survival rate below 20%. Accurate diagnostic tools predicting clinical outcomes and disease progression are desperately needed. We developed PathoNet, a novel deep-learning-based diagnostic software that automates immunohistochemistry scoring. PathoNet was uniquely designed with four steps: (1) formatting images into trainable tiles, (2) passing tiles through FilterNet, a convolutional neural network, and (3) ExpressNet, another convolutional neural network; and (4) aggregating tile scores to a final score. Instead of using packaged pre-trained models, we created our FilterNet and ExpressNet using the open-source PyTorch library, modeling after AlexNet architecture. PathoNet is currently optimized to score E-Cadherin (PathoNet-E-Cad), a biomarker that may predict EC progression and overall survival. Trained with 3072 tiles, PathoNet scores showed 85.62% tile-level concordance and 91.67% image-level concordance with pathologists, outperforming published automated immunohistochemistry scoring systems. We demonstrated the clinical potential of PathoNet-E-Cad by testing on 473 patient samples. The PathoNet-E-Cad score is associated with esophageal disease progression. Low PathoNet-E-Cad score is significantly correlated with better overall survival (p=0.043) and predicts optimal treatment outcomes of EC (p=0.027). More biomarkers are being integrated into PathoNet to further facilitate EC prognosis.

1 Introduction

1.1 Esophageal Cancer

Esophageal cancer is the eighth most common cancer and the sixth most common cause of cancer related deaths worldwide.[1] There are two major histopathological subtypes of esophageal cancer: esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC).[2] These two subtypes differ significantly in regards to epidemiological distribution, risk factors and clinical and prognostic relevance. EAC predominates in certain developed nations, for example, the United States. About 87% of all esophageal cancer cases globally are ESCC, with the highest incident rates seen in the Asian/Eastern countries.

Despite many advances in screening, diagnosis and treatment, the prognosis of ESCC is still poor: the 5-year survival rate for ESCC patients ranges from 10% to 20%.[3] The current optimal treatment option is neoadjuvant chemoradiation therapy with surgery (CRT), and other options include neoadjuvant chemotherapy, neoadjuvant radiation with surgery and stand-alone surgery. At present, clinical treatment decisions are based on tumor-node-metastases (TNM) staging; however, the clinical outcomes often display considerable variability in disease progression and survival. Better knowledge of patient prognosis and treatment prediction would

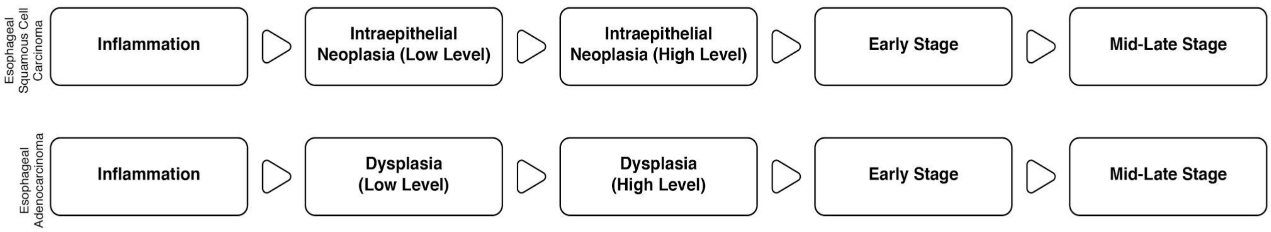
Esophageal Squamous Cell Carcinoma
Inflammation
Intraepithelial Neoplasia (Low Level)
Intraepithelial Neoplasia (High Level)
Early Stage
Mid-Late Stage
Esophageal Adenocarcinoma
Inflammation
Dysplasia (Low Level)
Dysplasia (High Level)
Early Stage
Mid-Late Stage

Figure 1. Esophageal cancer progression. Esophageal Squamous Cell Carcinoma is illustrated in the upper row, and Esophageal Adenocarcinoma in the lower row.

significantly help to facilitate personalized therapy decision-making for esophageal cancer patients.

Several emerging biomarkers evaluated by immunohistochemistry (IHC) have been reported as potential prognostic and predictive biomarkers, including E-cadherin (E-Cad), the Gli family of transcription factors and several molecules in signaling pathways such as Wnt, MAPK, and RAS signaling.[4] E-Cad is a 120-KD transmembrane calcium-dependent cell adhesion protein that has been implicated in cancer progression and metastasis.[5] Evidence to support E-Cad's prognostic value has been accumulated in spite of controversial and inconclusive reports in literature. The lack of a well-established and standardized scoring system may be one of the reasons contributing to the controversy. Here, we propose that an automated and standardized diagnostic/prognostic tool may be a promising and clinically practical solution.

1.2 Machine Learning in Medical Imaging

Machine learning is a rapidly expanding sub-field in medicine, with applications ranging from ailment diagnosis to biotechnological structure design.[6] In recent years, the relevance of machine learning in cancer diagnosis has been considered.[7] Previous attempts at classifying breast cancer and lung cancer in slide H&E images resulted in 70% and 79.7%[8] classification accuracies, respectively, using a convolutional neural network (ConvNet).[9] No studies using a similar ConvNet solution have been conducted on esophageal cancer, nor in developing automated and standardized diagnostic/prognostic tool.[10] We propose a novel application of machine learning technology to standardize IHC-stain intensity scoring for esophageal cancer.

1.3 Objectives

The current project aims to develop a novel deep learning based automated diagnostic/prognostic tool and provide proof-of-concept evidence for its clinical utility in indicating prognosis and predicting clinical response of current standard care CRT in order to aid personalized therapy decision-making for esophageal cancer patients.

2 Methodology

2.1 Esophageal Cancer Patient Samples

A total of 473 formalin-fixed, paraffin-embedded (FFPE) tissues were retrieved. Among all of these patients, 443 of them were originally diagnosed with either esophageal inflammation, intraepithelial neoplasia, or early stage and mid-late stage esophageal squamous carcinoma (ESCC). Additionally, 30 samples from esophageal adenocarcinoma (EAC) were collected. FFPE tissue blocks slides were cut into 5µm-thick sections. Written consent was obtained from each patient before specimen collection.

2.1.1 Immunohistochemistry Staining

Immunohistochemical staining (IHC) was performed following standard procedures. Briefly, FFPE slides were deparaffinized using xylene. Heat-mediated antigen retrieval was performed using a citrate buffer. Slides were stained by means of IHC for E-Cadherin (rabbit anti-human E-Cadherin; Cell Signaling) at a 1:200 dilution, EMX2 (rabbit anti-human EMX2, Pierce) at 1:400, and Gli2 (goat anti-human Gli2, Abcam) at 1:100. Antibody staining was visualized with DAB (Histostain Plus Broad Spectrum, Invitrogen) and hematoxylin counterstain (Fisher Scientific). Representative fields were photographed using an Olympus BX43 microscope (Olympus, Japan) and examined for positive nuclear staining.

2.1.2 Pathologist Labeling

The IHC staining of E-Cadherin was scored by a pathologist who was not aware of the corresponding clinical information. An IHC score was assigned by a combination of staining intensity (no staining = 0-, light yellow staining = 1+, yellowish brown staining = 2+, strong

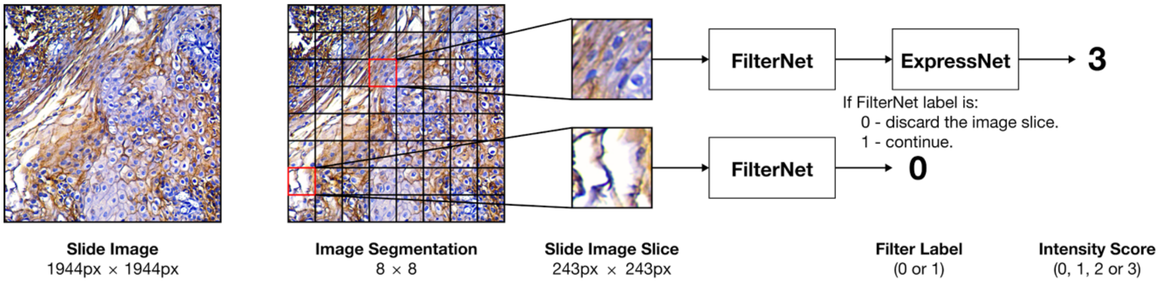

Figure 2. Flowchart of the PathoNet methodology. First, a slide image is segmented into 64 tiles. Each of these images is then passed into FilterNet, which determines if a particular tile contains cells or not. Tiles with cells are then passed into ExpressNet, which gives an intensity score for each tile. Final scores are calculated using the system described in Section 2.2.5, which is not shown in the diagram.

brown staining = 3+) and the percentage of positively stained cells. A score of 0- was defined as negative, and a score of 1+, 2+, or 3+ was defined as positive.

2.2 PathoNet

We have developed a methodology called PathoNet. PathoNet is designed with four main steps: 1) preprocessing and segmenting the slide images into tiles (image segmentation), 2) filtering out the individual tiles which do not contain cells (FilterNet), 3) assigning intensity scores for the remaining tiles (ExpressNet) and 4) using all individual intensity scores to produce an aggregate score representative of the whole image **(**final scoring systems).

2.2.1 Dataset

The dataset consists of E-Cadherin IHC images of 72 esophageal tissue samples at 10x magnification with a resolution of 1944x1944 pixels. Two training-testing setups have been evaluated, the Random setup and Balanced setup. The Radom training setup employs random assignment to assign the 72 slide images into a random training set of 48. The Balanced setup is curated to create a balanced training set of 48 images with a similar distribution of disease stages as the overall 473 patient cohort. The two setups use the same testing set of 12 images during performance comparison of ExpressNet. The testing set for the final scoring system is composed of 24 images that are not included in the Balanced training set.

2.2.2 Image Segmentation

Each of 72 slide images is segmented into 64 square tiles, creating an 8×8 grid of 243x243 pixel image tiles. There are 4,608 tiles useable for training and testing.[11] Pathologists manually labeled the 72 images at image level and the 4,608 tiles following features: 1) the overall score of the tissue sample at image level, 2) the individual scores of the 64 image tiles, 3) the percentage of area covered by cells in each image tile and 4) the percentage of the area covered by tumor cells in each image tile. To create training and testing sets for FilterNet and ExpressNet, the tiles are split at the image level, that is, all 64 individual tiles of one image are assigned together to either the training or testing set.

2.2.3 FilterNet

Individual tiles are filtered based on percentage of cell coverage. A deep convolutional neural network (deep ConvNet), called FilterNet, has been constructed to perform the filtering. FilterNet models the idea of stacking layers from AlexNet[12], consisting of two convolutional layers with the kernel sizes of 3x3.[13] More specifically, the convolutional layers are followed by a rectified linear unit (ReLU) and a MaxPooling operation, all leading into two fully connected layers that feed to the output. The specific model structure is design to have less layers than AlexNet to adopt with smaller number of training images compared to the ImageNet task of AlexNet. To train the FilterNet, each of 4,608 image tiles in the dataset are assigned a new label: 0 if the percentage of cell coverage is less than or equal to 50% and 1 if the percentage of cell coverage greater than 50%. Probabilities for each class are generated using a multinomial logistic regression (Softmax) operation, and the higher probability of the two classes is picked as the final prediction.

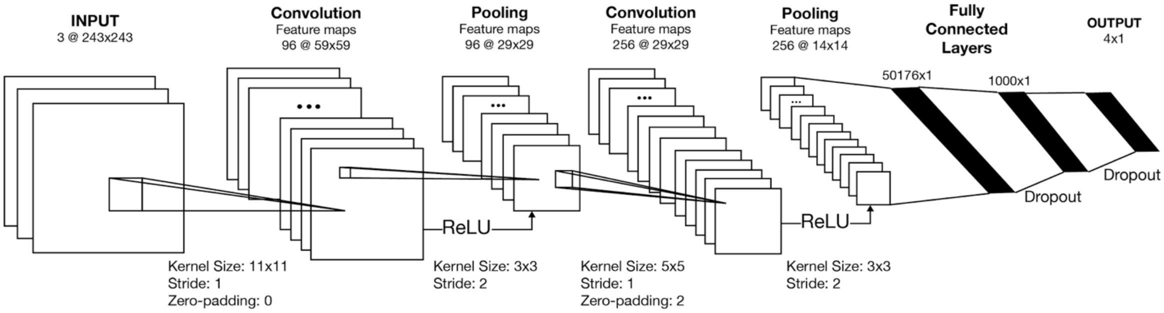

Figure 3. Flowchart of the ExpressNet architecture, which contains fewer convolutional layers compared to AlexNet.

2.2.4 ExpressNet

A second deep ConvNet, called ExpressNet, has been developed for the second classification of staining expression intensity on individual image tiles. This network's architecture (Figure 3) is also obtained from modifying the AlexNet, adding in dropout operations with a probability of 20% before each fully connected layer to prevent overfitting.[14] Similar to FilterNet, a softmax operation was performed to generate probabilities, and the highest probability out of the four categories—0- as low expression, 1+ and 2+ as intermediate expression and 3+ as high expression—was assigned as the prediction. To train the ExpressNet, individual tiles from the training cohort are passed in with intensity labels for either twenty or thirty epochs.

2.2.5 Final Scoring Systems

After individual tiles pass through FilterNet and ExpressNet, the staining intensity scores predicted by ExpressNet are aggregated to produce a final intensity score of the entire image using PathoNet's scoring algorithm. Three different scoring methodologies are evaluated: 1) weighted average scoring, 2) majority votes scoring and 3) modified majority votes scoring using a 30%-threshold pathologist rule.

$$p_{i,k} = \frac{exp(y_{i,k})}{\sum_{j=0}^{3} exp(y_{i,j})} \quad (1) \qquad \bar{x}_k = \frac{\sum_{i=0}^{63} \hat{y}_{i,k} p_{i,k}}{\sum_{i=0}^{63} p_{i,k}} \quad (2)$$

The weighted average scoring system first normalizes the output from ExpressNet by converting the largest of the four output energies $y_{i,k}$ into a probability $p_{i,k}$ that image tile i out of 64 total tiles belongs to that label k (Equation 1). This process is performed over all 64 tiles to find probabilities of $\hat{y_i}$ being the true label, calculating the weighted average $\bar{x}_k$ of the entire slide image as a continuous value ranging from 0-3 (Equation 2).

The majority votes system tallies individual predictions for tiles, and the label with the most tallies is chosen as the output.

The 30%-threshold system is designed by learning from how pathologists conventionally score the images. Built on top of the majority votes system, this system chooses a higher index (3+ > 2+ >1 + > 0-) as final output for the image when the higher index has at least 30% of votes, even if a lower index holds the most votes.

2.3 Statistical Analysis

Statistical analysis was performed in Microsoft Excel and OriginLab Origin 9.0. Machine learning performance was assessed with accuracy and F1 scores. TP, FP, FN and TN stand for true positive, false positive, false negative and true negative.

$$Precision = \frac{TP}{TP + FP} \quad (3) \qquad Recall = Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4) \qquad F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \quad (6)$$

Clinical significance was assessed by Kruskal-Wallis H test followed by a post-hoc Conover test, Pearson chi-square test and a Kaplan-Meier curve with a log rank analysis. We used an alpha value of 0.05 to determine statistical significance.

3 Results

3.1 E-Cadherin analysis in patient esophageal tissues

We systematically investigated protein expression levels of E-Cad, EMX2 and Gli2 as three potential progression and prognostic biomarkers in all 473 patients' specimens. On the progression of ESCC, histologic analysis showed 38 (8.6%) inflammation, 35 (7.9%) low grade intraepithelial neoplasia, 35 (7.9%) high grade intraepithelial neoplasia, 212 (47.9%) early stage ESCC, and 123 (27.8%) mid-late stage ESCC (Table 1). An additional 30 EAC samples were also included in the analysis. Protein expression levels were characterized by IHC (representative positive and negative E-Cadherin expressions in different esophageal disease stages are shown in Figure 4) and scored on a scale of 0-3 (negative, mild, moderate and strong positive) (Figure 5). The score was determined by a pathologist and recorded as "Pathologist Score." The images serve as our unique dataset for PathoNet development.

3.2 Development of PathoNet E-cadherin Score

PathoNet process has four main steps: 1) preprocessing and segmenting the slide images into tiles (image segmentation), 2) filtering out the individual tiles which do not contain cells (FilterNet), 3) assigning intensity scores for the remaining tiles (ExpressNet) and 4) using all individual intensity scores to produce an aggregate score representative of the whole image **(**final scoring systems). The performance of individual step and overall process was evaluated systematically.

Lesion Type	Number	Percent
Inflammation	38	8.58%
Intraepithelial Neoplasia		
Low Grade (LIN)	35	7.90%
High Grade (HIN)	35	7.90%
Esophageal Squamous Cell Carcinoma		
Early Stage	212	47.86%
Mid-Late Stage	123	27.77%
Total	443	

Table 1. Patient Information

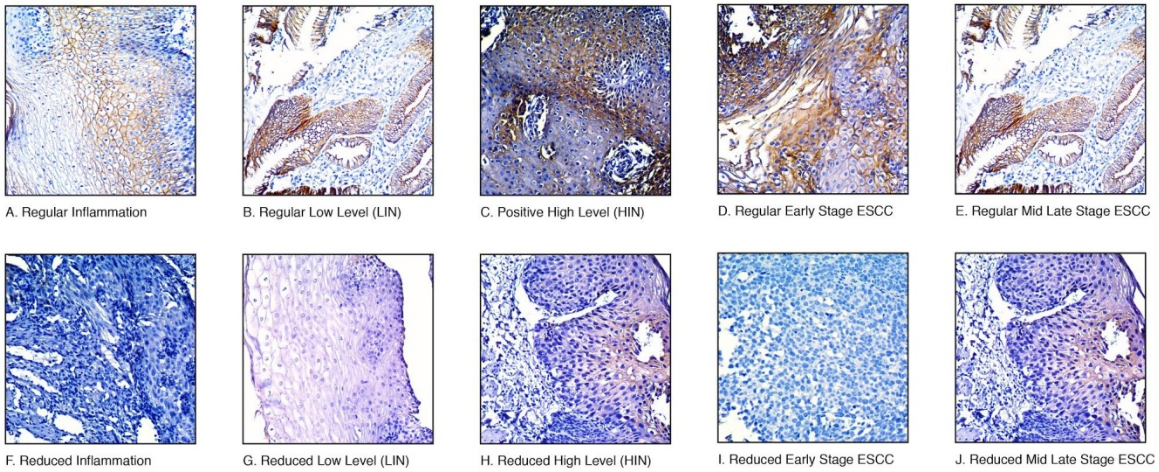
A. Regular Inflammation
B. Regular Low Level (LIN)
C. Positive High Level (HIN)
D. Regular Early Stage ESCC
E. Regular Mid Late Stage ESCC
F. Reduced Inflammation
G. Reduced Low Level (LIN)
H. Reduced High Level (HIN)
I. Reduced Early Stage ESCC
J. Reduced Mid Late Stage ESCC

Figure 4. Representative "+" and "-" IHC staining of E-Cad in different esophageal disease stages. A-E) Positive (+) staining of E-Cad. F-J) Negative (-) staining of E-Cad. Disease stages are labeled under each image.

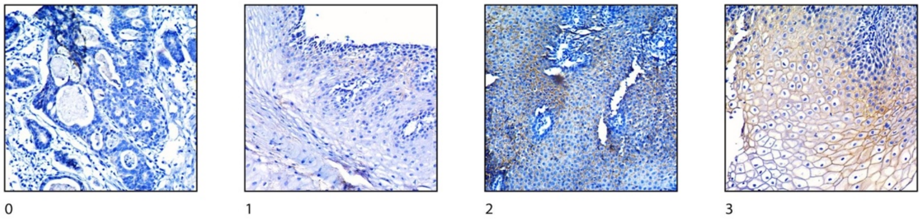

Figure 5. Representative scores of IHC staining of E-Cad. From left to right, the score is 0-, 1+, 2+, 3+ (negative, mild, moderate and strong positive), respectively.

From our unique dataset described above, we chose images of tissue samples from 72 patients to develop PathoNet.

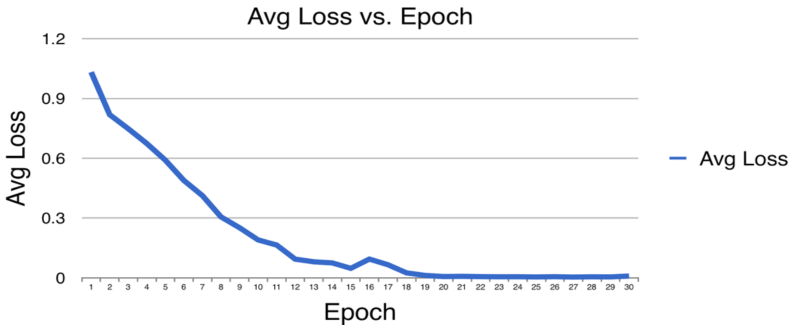

Figure 6. The average loss of ExpressNet collected over 30 epochs. Note that the loss flatlines after 20 epochs, meaning that classification accuracy no longer increases after that point.

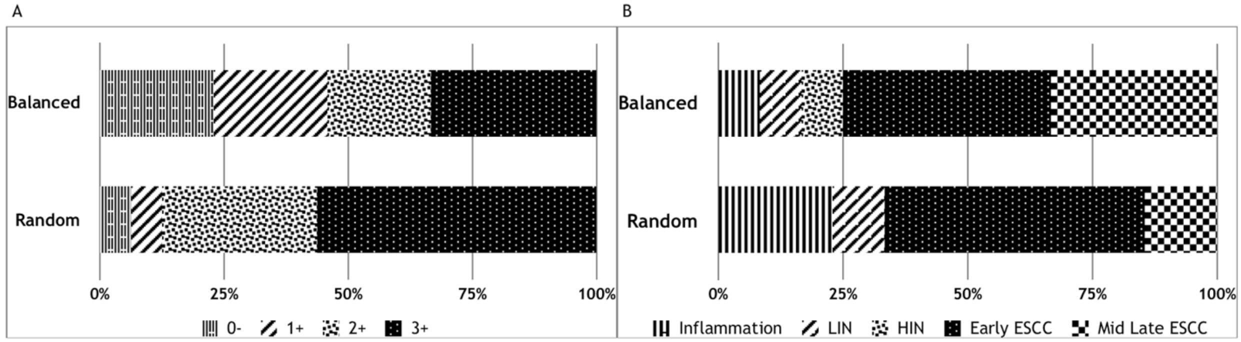

Figure 7. ExpressNet training set sample distribution in Balanced setup and Random setup. The training set in Random setup was obtained by random assailment and that Balance setup by manual curation. A) Sample distribution by image level IHC score. B) Sample distribution by disease stages. Balanced setup distribution mimics overall patient sample distribution summarized in Table 1.

3.2.1 FilterNet Performance

Trained on individual image tiles, the FilterNet achieved a 94.53% classification accuracy in distinguishing between tiles that had a large percentage of the area covered in cells and tiles that had little to no cell coverage.

3.2.2 Overfitting of ConvNets

The ExpressNet was trained for up to 30 epochs, and our results suggested that the model trained for only 20 epochs produced a higher classification accuracy on the testing set. The disparity between more epochs of training and classification accuracy is due to overfitting, which occurs when the weights of the neural network are too highly specified for the training dataset.

	F1 Score				
	0-	1+	2+	3+	Overall Accuracy
Random Training Set	86.80%	47.00%	45.80%	85.50%	70.30%
Balanced Training Set	87.40%	44.30%	55.40%	91.60%	85.62%

Table 2. Classification performance of the ExpressNet with the Random training set and the Balanced training set. The F1 score, a measure of precision and recall, was calculated for each model at each level.

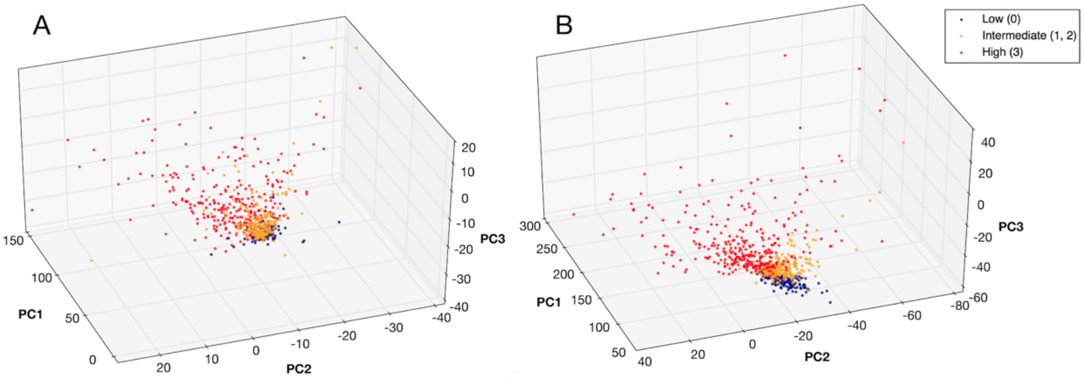

Figure 8. Principal components analysis of machine learned features of A) Random setup and B) Balanced setup. Balanced setup has a better separation of features than Random setup. In this case, labels 1+ and 2+ are combined as "intermediate" to show clearer divisions.

We decided to use the ExpressNet for 20 epochs, since further training no longer reduced loss (Figure 6).

3.2.3 ExpressNet Performance

ExpressNet generates tile level classification of 0-, 1+, 2+ and 3+ for E-Cad expression scores. We compared the ExpressNet performance of the Random setup (training set by random assailment) and Balance setup (training set by manual curation) with training set distribution illustrated in Figure 7. The overall accuracy of 768 tiles from 12 images was 85.62% of Balanced setup, rising from 70.30% of Random setup (Table 2).

To have a better understanding of advantages that the different datasets provided, we performed principal component analysis (PCA) to display the features extracted by the ExpressNet (Figure 8). The PCA charts show that the ExpressNet trained on the curated dataset

(Balanced setup) had better separation of features than the ExpressNet trained on the randomized dataset (Random setup) (Figure 8).

3.2.4 PathoNet Final Scoring System Performance

The final scoring system aggregates 64 individual tile classifications to produce an image level score. We compared the performance of three scoring systems 1) weighted average, 2) majority votes and 3) a 30%-threshold system by testing them with 24 images.

The weighted average system produced a continuous value from 0-3. The root mean squared error (RMSE) of the weighted averages compared to the true integer labels was 0.5685. To calculate accuracy, we found the label that was closest to the continuous value by rounding it to the nearest integer. Then, comparing those rounded predictions to the true labels yielded a 66.67% accuracy where 16 out of 24 testing images were correctly classified.

Unlike the weighted average system, the majority votes system returned integer outputs for predictions, where 21 out of 24 testing images were correctly classified, yielding an 87.5% accuracy.

The 30%-threshold system is based on the manual scoring rule of pathologists, achieving the best performance with an accuracy of 91.67%, where 22 out of 24 testing images were correctly classified.

Therefore, we finalized the PathoNet final scoring system with the 30%-threshold method. The PathoNet E-Cad method, with an overall accuracy of 91.67%, outperforms previous literature[15].

3.3 PathoNet E-Cad score serves as a promising progression and prognostic biomarker in esophageal cancer

3.3.1 PathoNet E-Cad score serves as a promising progression biomarker in ESCC

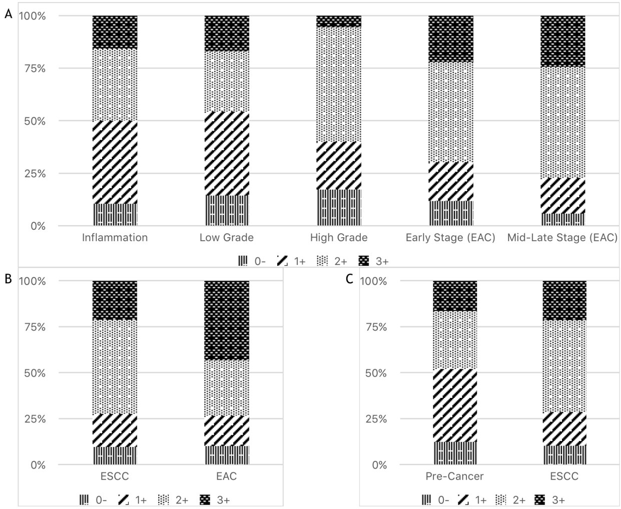

Figure 9. The distribution of PathoNet E-Cad score across A) different disease stages, B) two subtypes as squamous-cell carcinoma (ESCC) and adenocarcinoma (EAC) and C) pre-cancer stages and cancer.

	Kruskal-Wallis H test	
Adjusted H	**d.f.**	**p-value**
11.248	4	0.024
	Pairwise Post-hoc Conover Test	
Pathologic Stage (n)	Pathologic Stage (n)	Conover Test
Mid-late ESCC (123)	Early ESCC(212)	P>0.05
	HIN(35)	p<0.01*
	LIN(35)	p<0.01*
	Inflammation(38)	p<0.05*
Early ESCC (212)	HIN(35)	p<0.05*
	LIN(35)	p<0.05*
	Inflammation(38)	P>0.05
HIN (35)	LIN(35)	P>0.05
	Inflammation(38)	P>0.05
LIN (35)	Inflammation(38)	P>0.05

Table 3. E-Cadherin analysis of different disease stages in esophageal cancer. Kruskal-Wallis H test and post-hoc Conover test were employed to access distribution difference. P<0.05 was considered significant and labeled with *.

To evaluate if PathoNet E-Cad is a promising biomarker in esophageal cancer, we applied the newly developed tool PathoNet to determine E-Cad expression levels in all patient samples.[16] The PathoNet E-Cad score presented different distribution of E-Cad protein expression at different disease stages of ESCC progression (Figure 9). In order to test if PathoNet E-Cad varies significantly at different disease stages, we employed the non-parametric Kruskal-Wallis H test, the result of which revealed the expression difference was significant between individual groups (p=0.024, Table 3). Subsequently, Conover tests were performed for pairwise multiple comparisons to discern which of many possible sample pairs were significantly different (Table 3). Significant difference was observed between two ESCC stages (mid-late and early)

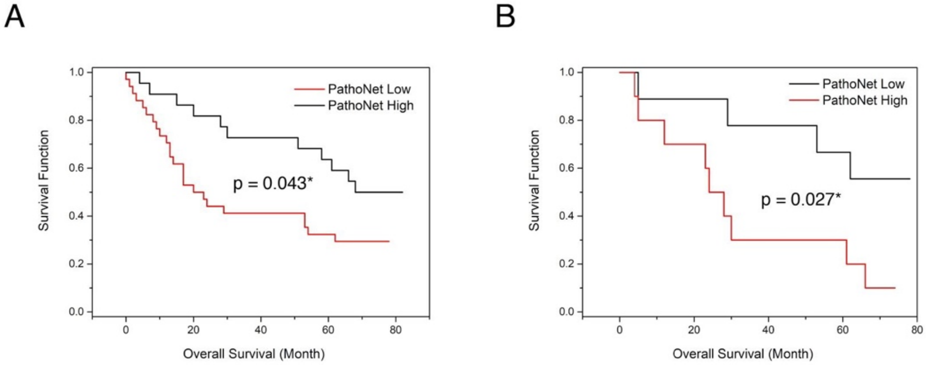

Figure 10. PathoNet E-Cad score may serve as a prognostic marker. Kaplan-Meier Curves for the overall survival of A) all ESCC patients and B) for patients treated with CRT. Log rank test was performed to generate p values, and P<0.05 was considered significant and labeled with *.

and three pre-cancer stages (HIN, LIN and inflammation). respectively, supporting the hypothesis that E-cad expression changes during ESCC progression and that PathoNet E-Cad may serve as a progression biomarker for ESCC.[17] Consistently, a Pearson Chi-square test conducted between ESCC and pre-cancer stages showed a significant difference (P=0.0004 P<0.001), further suggesting PathoNet E-Cad as a good biomarker to differentiate ESCC from pre-cancer stages.[18] No significant changes were observed between HIN and LIN, or between mid-late and early ESCC, indicating that PathoNet E-Cad expression changes were not a stand-alone marker to distinguish stages within intraepithelial neoplasia or ESCC as shown in Table 3. Further analysis revealed that PathoNet E-Cad exhibited distinct distributions between ESCC and EAC (P=1.3E-08, P<0.001) (Figure 9), consistent with the consensus that the two subtypes of esophageal cancer significantly differ in molecular mechanisms.

3.3.2 PathoNet E-Cad serves as a promising prognostic biomarker in ESCC

To investigate if PathoNet E-Cad is a prognostic biomarker in ESCC, we analyzed the correlation between PathoNet E-Cad scores and patients' overall survival as well as their treatment strategies. Data were collected from 56 cases of esophageal patients that had undergone surgery with neoadjuvant chemoradiation therapy (CRT) (n=19), surgery with neoadjuvant radiation therapy (n=10), neoadjuvant chemotherapy (n=17) or surgery alone

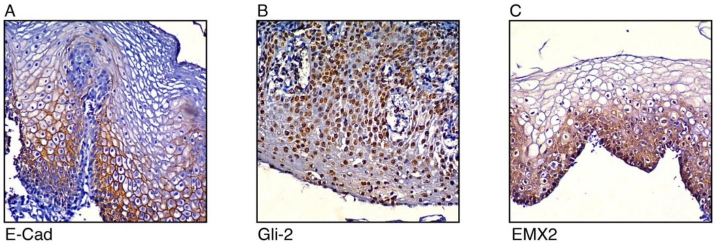

Figure 11. Biomarkers evaluated by IHC and ready for PathoNet modeling. A) E-Cad, B) Gli-2 and C) EMX-2 showed different staining patterns.

(n=10). A high PathoNet E-Cad score is significantly associated with improved overall survival in all 56 patients when analyzed with a Kaplan-Meier Curve to generate a log rank p-value of 0.043 (Figure 10A). The median overall survival was 36.6 months (95% confidence interval 26.3-46.9 months) vs 59.5 months (95% confidence interval 47.9-71.2 months). Among all treatment strategies, CRT is currently the most optimal method to manage resectable ESCC. Notably, a high PathoNet E-Cad score is significantly associated with improved overall survival in patients with CRT (Figure 10B) (p=0.027). The median overall survival was 32.7 months (95% confidence interval 17.7-47.6 months) vs 59.9 months (95% confidence interval 43.5-76.3 months). The result strongly suggested PathoNet E-Cad as a prognostic biomarker as well as predictive marker for ESCC patients with CRT treatment. We observed a separation of OS survival curves in neochemotherapy and neoradiation groups; however, the difference was not very significant ($p > 0.05$). No difference was found in patients who had undergone stand-alone surgery.

3.3.3 EMX2 may augment prognostic value of PathoNet in ESCC

We hypothesize that the integration of multiple biomarkers into PathoNet which is further augmented with an algorithm based final scoring system would significantly improve the performance of PathoNet as a prognostic and diagnostic tool and thereby its clinical utility. Several categories of molecules are associated with esophageal cancer progression and prognosis, such as those in the Wnt and Hedgehog signaling pathways. We investigated the expression of new biomarkers EMX2, a homeo-domain containing transcription regulator, and Gli2, downstream transcription factor of Hedgehog signaling in 473 patients samples (Figure 11).[19] Preliminary data based on pathologist score showed that EMX2 was associated with overall survival (OS) in a limited number of ESCC patients treated with neoadjuvant radiotherapy (n=10, p=0.028*). Efforts have been made to train PathoNet with EMX2 and Gli2.

4 Discussions and Future Directions

In this project, we have developed a novel deep learning based diagnostic tool called PathoNet for predicting progression and prognosis of esophageal cancer patients and established proof of concept evidence for its potential clinical application.

The benefit of using computer-aided diagnosis for tissue biomarkers in pathology has not been well-established. Recently, research has provided an increasing amount of evidence to suggest that the capability of deep learning to achieve complex pattern recognition could lead to a new generation of computer-aided diagnoses. Major efforts in the field of pathology has been made in classification of tumor vs. non-tumor tissues, however the biomarker score based diagnosis, such as HER2 IHC evaluation and E-Cad in the current project, has very limited progress. Compared to recent machine learning attempts to classify IHC-staining intensities, our results show an improvement with a classification accuracy on an individual tile basis of 85.62% compared to 83% in Vandenberghe *et al.*[15]Our results showed that F1 scores for classifying individual tiles labeled 1+ or 2+ were low compared to tiles labeled 0- or 3+ (Table 2). Our discordance analysis (Figure 12) showed this same trend, where the larger bubbles on the (0, 0) and (3, 3) points showed high concordance for tiles labeled 0- and 3+. The smaller bubbles in between showed discordance for tiles labeled 1 and 2. In context, this means that the network is able to distinguish between negative expression (0-) and strong positive expression (3+) at tile

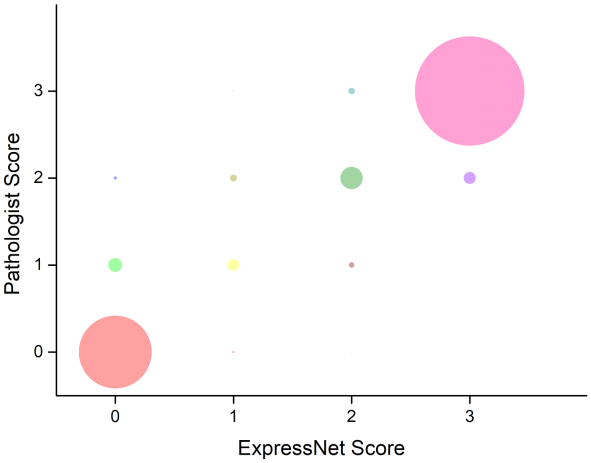

Figure 12. Discordant analysis of ExpressNet for tile level scoring. ExpressNet scores at x axis are compared with Pathologist scores at y axis of 0-, 1+, 2+ and 3+ (0-3 in the chart). Bubble position shows the comparison of ExpressNet and Pathologist scores. The size of bubbles corresponds to the number of samples that fall into each category.

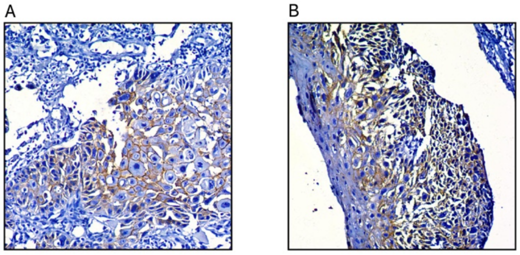

Figure 13. The discordant images from overall performance testing. Image A was classified as 0- by the ExpressNet, and its pathologist score is 2+. Image B was classified as 3+ by the ExpressNet, and its pathologist score is 2+.

level more easily than it can distinguish low to moderate positive expression (1+, 2+). We hypothesize that the reason for this lies in our skewed dataset, since a majority of tiles is labeled either a 0 or a 3. More specifically for our testing dataset, 28.2% of individual tiles was labeled as 0 and 42.5% was labeled as 3, leaving only 11.7% labeled as 1 and 16.7% labeled as 2. In addition, this trend is consistent with published data where intensity based scoring, i.e. distinguishing 1+ and 2+, is an area ConvNet in general needs improvement[15]. Historically, pathologists manually label whole slide images. As an automatic and computer-aided approach, PathoNet can be applied to standardize IHC scoring and remove the subjective and error-prone procedure by pathologists.[20] Potts *et al.*[21] states that discordant scoring of staining intensity between pathologists is largely due to tumor heterogeneity. It was found that tissue samples with scores +1 and +2 have the highest levels of tumor-level heterogeneity, consistent with the results of the classifier. Although this automated system still requires input data from digital microscopy images, which could vary depending on where the data were retrieved from[22], this study did not examine slide samples retrieved from different labs. Final results from testing the PathoNet methodology showed 2 out of 24 testing images, shown in Figure 13, as incorrectly classified. Reanalysis of these two discordant cases help us improve our algorithm to further distinguish tumor cells from non-tumor cells. To improve the technical and clinical performance of PathoNet as a prognostic and diagnostic tool, we have initiated the following efforts. First, we look to provide the PathoNet model with more training data with more labeled features, which would increase the classification accuracy for individual tiles. Second, we will further define a final scoring system by integrating more features. Third, we have started to incorporate more biomarkers, such as EMX2 and Gli2, as biomarker panels provide more clinical and mechanistic information. Finally, we will generate a prognostic/diagnostic score by integrating individual biomarker scores with a new scoring algorithm. We believe deep machine learning based diagnostic software holds the promise to provide a clinically useful tool to address needs such as esophageal cancer prognosis as studied in this project.

5 Conclusions

PathoNet is, in this study, optimized to score E-Cadherin, a biomarker that may predict EC progression and overall survival. Trained with 3,072 tiles, PathoNet E-Cad scores showed tile-level concordance with pathologists of 85.62% with 1536 tiles and image-level concordance of 91.67%, outperforming published automated immunohistochemistry scoring systems. We also demonstrated the clinical potential of PathoNet E-Cad by testing on 473 patient samples. The score is associated with esophageal disease progression. Low scores are significantly correlated with better overall survival (p=0.043) and predict optimal treatment outcomes of esophageal cancer (p=0.027). In the future, more biomarkers will be integrated into PathoNet to further facilitate esophageal cancer prognosis.

6 References

1 Enzinger, P. C. and Mayer, R. J., Esophageal cancer. *N Engl J Med* 349 (23), 2241 (2003).

2 Hameeteman, W., Tytgat, G. N., Houthoff, H. J., and van den Tweel, J. G., Barrett's esophagus: development of dysplasia and adenocarcinoma. *Gastroenterology* 96 (5 Pt 1), 1249 (1989).

3 Wang, K. K. and Sampliner, R. E., Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus. *Am J Gastroenterol* 103 (3), 788 (2008).

4 Xu, X. L. et al., The impact of E-cadherin expression on the prognosis of esophageal cancer: a meta-analysis. *Dis Esophagus* 27 (1), 79.

5 Hirata, D. et al., Involvement of epithelial cell transforming sequence-2 oncoantigen in lung and esophageal cancer progression. *Clin Cancer Res* 15 (1), 256 (2009).

6 Granter, S. R., Beck, A. H., and Papke, D. J., Jr., AlphaGo, Deep Learning, and the Future of the Human Microscopist. *Arch Pathol Lab Med* 141 (5), 619.

7 Laak, J. A., Pahlplatz, M. M., Hanselaar, A. G., & Wilde, P. C., Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy. *Cytometry* 39 (4), 275 (2000).

8 Sun, W., Zheng, B., & Qian, W., Computer aided lung cancer diagnosis with deep learning algorithms. *Medical Imaging 2016: Computer-Aided Diagnosis* (2016).

9 Bejnordi, B. E. et al., Stain Specific Standardization of Whole-Slide Histopathological Images. *IEEE Trans Med Imaging* 35 (2), 404 (2015).

10 Sompuram, S. R. et al., Standardizing Immunohistochemistry: A New Reference Control for Detecting Staining Problems. *J Histochem Cytochem* 63 (9), 681.

11 Hou, L. et al., Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016, 2424.

12 Krizhevsky, A., Sutskever, I., & Hinton, G. E, ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60 (6), 84 (2012).

13 Dumoulin, V., & Visin, F. , A guide to convolution arithmetic for deep learning. . *eprint arXiv:1603.07285. Retrieved September 24, 2017.* (2016).

14 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 1929 (2014).

15 Vandenberghe, M. E. et al., Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci Rep* 7, 45938.

16 Okumura, H. et al., Biomarkers for predicting the response of esophageal squamous cell carcinoma to neoadjuvant chemoradiation therapy. *Surg Today* 44 (3), 421.

17 Tang, N. N. et al., HIF-1alpha induces VE-cadherin expression and modulates vasculogenic mimicry in esophageal carcinoma cells. *World J Gastroenterol* 20 (47), 17894.

18 Wang, C. et al., Immunohistochemical prognostic markers of esophageal squamous cell carcinoma: a systematic review. *Chin J Cancer* 36 (1), 65.

19 Zhu, W. et al., Correlation of hedgehog signal activation with chemoradiotherapy sensitivity and survival in esophageal squamous cell carcinomas. *Jpn J Clin Oncol* 41 (3), 386.

20 Granter, S. R., Beck, A. H., and Papke, D. J., Jr., Straw Men, Deep Learning, and the Future of the Human Microscopist: Response to "Artificial Intelligence and the Pathologist: Future Frenemies?" *Arch Pathol Lab Med* 141 (5), 619.

21 Potts, S. J. et al., Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. *Lab Invest* 92 (9), 1342.

22 Szolovits, P., Patil, R. S., and Schwartz, W. B., Artificial intelligence in medical diagnosis. *Ann Intern Med* 108 (1), 80 (1988).